STAT 300 Chapter 6
Point Estimation

Zepu Zhang
March 22, 2012

1 Point estimate, point estimator

Suppose we want to "estimate" some unknown quantitative property, say mean, variance, upper quartile, whatever, of the population.

Usually we're interested in estimating the "parameter(s)" that we use to describe the distribution. For example, if the population distribution is normal, we want to estimate mean (μ) and variance (σ^2); if the population distribution is binomial, we want to estimate the success rate (p); if the population distribution is exponential, we want to estimate the parameter λ (reciprocal of mean).

However, the concept of "estimation" is not restricted to such "characteristic properties"; it can be about any quantitative property (e.g. the 93th percentile of the population distribution).

For generality, let's use θ to denote the population parameter to be estimated.

Definition Point estimate and point estimator. A point estimate is a single number that is regarded a sensible value for θ. An estimate is necessarily a function of the sample data. This function (the formula; before plugging in the actual data) is called the point estimator of θ.

We first have an estimator (that is, what function of the sample data we will use to estimate θ); then we plug in the actual sample data to get a specific value—the estimate. Both the estimator and the estimate are denoted by $\hat{\theta}$. (We may also choose to use a different symbol for the estimate.)

Note A "point estimator" is in contrast to an "interval estimator" (or "confidence interval"), to be discussed later.

A point estimator is a random variable because it is a

function of a random sample. The distribution of this random variable (realized in in repeated sampling) is called its sampling distribution. As with any random variable we want to examine its distribution.

There are multiple functions of the sample data that may appear as reasonable estimators of θ. This section is about the criteria of a "good" estimator.

Example Ex. 6.1 in Chap. 6.1. Estimating success rate, p, of a binomial distribution.

Example Ex. 6.2 in Chap. 6.1. Estimating population mean μ. Candidates: sample mean, sample median, trimmed mean, middle point of extremes, etc.

Example Estimating population variance, σ^2. Candidates: $\frac{1}{n-1}\sum(X_i-\overline{X})^2$ (the sample variance) and $\frac{1}{n}\sum(X_i-\overline{X})^2$ (something slightly different from the sample variance).

2 Criteria for "good" estimators

2.1 Criterion 1: unbiasedness

We like an estimator to be unbiased, that is, $E(\hat{\theta}) = \theta$.

Example Estimating p of binomial: $\hat{p} = X/n$ is unbiased (because $E(X/n) = E(X)/n = (np)/n = p$).

Example Estimating μ: sample mean $\overline{X}$ is unbiased (because $E(\overline{X}) = \mu$).

Example Estimating σ^2: $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$ is unbiased (but $\frac{1}{n}\sum(X_i - \overline{X})^2$).

Note $S = \sqrt{S^2}$ is the usually adopted estimator for σ. It is biased. $E(S) = E(\sqrt{S^2}) \neq \sqrt{E(S^2)} = \sqrt{\sigma^2} = \sigma$.

Now we have one criterion. Unfortunately this will not settle things yet, because for a specific θ there could be more than one unbiased estimator. For example, the proposition on page 233 (7th ed) or page 246 (8th ed).

2.2 Criterion 2: small variance

Among all unbiased estimators of θ, the one whose sampling variance is the smallest is called the minimum variance unbiased estimator, **MVUE**.

Note For a particular θ, if we know of a MVUE for it, we usually use it. For some problems, we may not know whether a MVUE exists or what it is. (See MLE below.)

> Sometimes an estimator is called "best"; that often means MVUE. (Usually nothing is absolutely "best" in every sense. Find out what is meant in the context and forget about "best".)

Example For normal distribution, sample mean $\overline{X}$ is the MVUE for μ.

Note $\overline{X}$ is always an unbiased estimator for μ, but it is not necessarily MVUE. For normal, it is. Ex. 6.7 in Chap. 6.1 gives examples where it is not.

> When presenting an estimator, we need to report on its bias (but we usually strive to use unbiased estimators) and variance. The former provides a measure of "correctness" whereas the latter, "precision".

As a measure of precision, we often report the standard deviation of $\hat{\theta}$, denoted by $\sigma_{\hat{\theta}}$. This is called the standard error. The standard error is typically connected to the unknown population distribution, in particular population variance (or standard deviation), and other population properties. In addition, $\sigma_{\hat{\theta}}$ is (perhaps) always related to sample size (n): $\sigma_{\hat{\theta}}$ decreases as n increases.

In order to report $\sigma_{\hat{\theta}}$, we often need to use estimates of population properties that are needed but unknown, that is, we often provide an estimate of $\sigma_{\hat{\theta}}$ (which may be denoted by $\hat{\sigma}_{\hat{\theta}}$).

Example Ex. 6.9 in Chap. 6.1.

Example Ex. 6.10 in Chap. 6.1.

(Skip the material on bootstrap.)

3 Maximum Likelihood Estimation (MLE)

We've learned some criteria for "good" estimators and examined several particular estimators (and judged whether they are biased, unbiased, good, bad, etc.). Now comes a recipe for actually finding an estimator (without relying on others to tell us to check this form, that form, etc.).

Definition **Likelihood function**: (joint) pdf or pmf of the sample data, viewed as a function of the parameter(s) θ.

A customary way to write it is $L(\theta; x_1, \ldots, x_n)$.

Note The value of $L(\theta;\, x_1, \ldots, x_n)$ is just the joint density function of $x_1, \ldots, x_n$. BUT, here the data $x_1, \ldots, x_n$ are fixed, whereas the parameter θ is viewed as a variable. One may try different values of θ to get different values of L.

Definition **MLE**: find the value of θ that maximizes the likelihood function; take that value as the estimate of θ; denote the estimate by $\hat{\theta}$. Mathematically, this is written as

$$\hat{\theta} = \operatorname*{argmax}_{\theta} L(\theta;\, x_1, \ldots, x_n)$$

Suppose the density (or mass) of X_i is $f(x_i; \theta)$, then because the sample is usually iid, the likelihood is typically

$$L(\theta; x_1, \ldots, x_n) = \prod_i f(x_i; \theta).$$

(The joint pdf or pmf of independent variables is equal to the product of the pdf or pmf of each individual variable.)

Remark

1. We need to know the pdf or pmf in order to proceed with MLE.
2. The likelihood function (joint pdf or pmf) is usually a big product, because the sample is usually independent.
3. To find the $\hat{\theta}$ that maximizes $L(\theta)$, it's usually more convenient to maximize $\log L(\theta)$ (the **log likelihood**).
4. Procedure for finding MLE:

(a) Write out the log likelihood function.

(b) Take derivative of $\log L(\theta)$ w.r.t. θ.

(c) Set the derivative to 0 and solve for θ (using calculus or numerical optimization algorithms), denoting the solution by $\hat{\theta}$.

(d) Check that the $\hat{\theta}$ just found is indeed a maximizer (instead of a minimizer) of $\log L(\theta)$.

If the unknown θ is a vector, we need to set to 0 the derivative of $\log L$ w.r.t. each element of θ, and solve an equation system.

Example Ex. 6.15 in Chap. 6.2.

Example Ex. 6.17 in Chap. 6.2.

Example 6.17 tells us that a MLE may be a biased one. But it can't be too bad, because—

Proposition **Large sample behavior of MLE**: when sample size n is large, mle $\hat{\theta}$ is approximately the MVUE of θ, under very general conditions and for any parameter θ.

Proposition **Invariance of MLE**: suppose $\hat{\theta}$ is the MLE for θ, then $h(\hat{\theta})$ is the MLE for $h(\theta)$.

MLE is a great thing mainly because of the two properties provided by these propositions. Of course, it's also very nice that finding MLE is a routine procedure.

Example Ex. 6.20 in Chap. 6.2.